On an adaptive version of the Metropolis-Hastings algorithm with independent proposal distribution

J. Gåsemyr

Abstract

In this paper we present a general formulation of an algorithm, the adaptive independent chain (AIC), that was introduced in a special context in Gåsemyr, Natvig and Sørensen (2000). The algorithm aims at producing samples from a specific target distribution Π, and is an adaptive, non-Markovian version of the Metropolis-Hastings independent chain (IC), see Hastings (1970), Tierney (1994). We show that under certain conditions, the algorithm produces an exact sample from Π in a finite number of iterations (with probability 1), and hence that it converges to Π. We also study features such as acceptance rate and autocovariance, and argue heuristically for the profitability of the adaptive procedure. We also study the asymptotic efficiency compared to that of rejection sampling, and indicate the relationship to importance sampling. A modified version of the AIC, the componentwise AIC (CAIC) is also introduced.

Key words: coupling, exact sampling, importance sampling, independent chain, Markov Chain Monte Carlo, rejection sampling, Bayesian hazard rate estimation.

1 Introduction

Let Π be a distribution on a space $\mathcal{X}$. We assume it is too complicated to draw a sample from Π directly. We also assume that it is impossible to compute $E_\Pi(h(X))$ analytically, where h is a function on $\mathcal{X}$. Rejection sampling, importance sampling and sampling importance resampling (SIR) are techniques for using samples from a different distribution P to obtain such samples or estimates. Another such method is to use P as a proposal distribution for a Metropolis-Hastings algorithm, independent of the current state of the Markov chain that is generated. Such a Markov chain is referred to as an independent chain (IC) in the terminology of Tierney (1994), while Liu (1996) refers to the algorithm as Metropolized independent sampling. A reasonable performance of these procedures depends on P not being too far from Π in some sense. If it is impossible to make a good guess at Π by an easily simulated P, other Markov chain methods, such as the Gibbs sampler or versions of the Metropolis-Hastings algorithm based on local moves around the current state, may be the solution. These Markov chain methods have the disadvantage that they often converge slowly, or at least that diagnosis of the convergence is difficult.

The problem of poor match between Π and P for an independent chain may in some cases be solved by letting P change adaptively during the running of the chain. We may refer

to this as an adaptive independent chain (AIC). This falls within the general framework of Holden (2000), where many references on adaptive algorithms can be found. A special case of the AIC is presented in Gåsemyr, Natvig and Sørensen (2000) along with two algorithms based on a similar idea, changing the proposal distribution as likelihoods from independent data sets are sequentially incorporated in a posterior distribution. In the present paper we focus on the adaptive aspect, but plan to return to applications within a sequential framework, such as state space models, in future work. In Gåsemyr, Natvig and Sørensen (2000) both a parametric version (PAIC) and a non parametric one (NPAIC) of the AIC are formulated, and the parametric one is shown to perform quite well.

The advantage of the AIC compared to the Markov chain methods making dependent moves, is that the nice convergence properties that are established for IC can be generalised. In fact, one may even obtain exact samples from Π using this method. The problem of obtaining exact samples when using MCMC techniques has been given much attention since the breakthrough made by Propp and Wilson (1996), who used the technique of coupling from the past to obtain exact samples. A very simple fact which seems to have been overlooked so far, is that one can obtain exact samples by forward coupling in the case of IC. This can be generalized to AIC. Compared to rejection sampling, importance sampling and IC, the AIC has the advantage that it is not so vulnerable to the choice of proposal distribution. Intermediate samples generated by means of a bad proposal do contain information that can be used to improve the proposal.

The notion of AIC is introduced more precisely in section 2 of this paper, and the mentioned results on convergence rate and exact sampling are stated and proved. We also compute acceptance probability and autocovariances under stationarity. In section 3 we present a general framework suitable for construction of AIC algorithms, and argue heuristically that the adaptive strategy should be beneficial and converge. We present as an example the model type discussed in Gåsemyr, Natvig and Sørensen (2000). We also discuss how to determine the burn in. In section 4 we discuss the relationship of the IC to rejection sampling and compare efficiencies of the IC and the AIC with that of rejection sampling asymptotically. For the IC, this has been done in the discrete case by Liu (1996) by means of spectral theory. Our discussion, covering also the continuous case, is based on purely probabilistic arguments. This argument in itself provides insight into the relationship between the algorithms, and shows that the best bound on the relative efficiency that is generally applicable, does in fact represent a very extreme worst case. Also, the relationship to importance sampling is discussed briefly.

Section 5 is more speculative in nature. The expected behaviour of the AIC as the dimension of the problem increases is considered, and some ideas concerning the construction of proposal distributions are presented. The ideas are demonstrated through an example referring to Arjas and Gasbarra (1996). A variant of our suggested algorithms called CAIC (componentwise adaptive independent chain), which may be considered as an adaptive approximation to the Gibbs sampler, is also presented. In appendix A we give sufficient conditions for the validity of our main theorem (Theorem 1), while appendix B gives a theoretical result that is not directly applicable to the AIC, but sheds more light on the behaviour of the algorithm, and may form the basis for more elaborate versions.

2 Definition and basic properties

Let $\{P_\phi\}$ be a family of probability distributions on the subspace $\mathcal{X}$ of R^n, with the parameter ϕ ranging in a subset Φ of R^d for some integer $d \geq 1$. Let p_ϕ be the density of P_ϕ, assumed to be continuous. We denote by π the density of the target distribution Π that we want to sample from, also assumed continuous. To specify an AIC we must specify a sequence length K and a function $\hat{\phi}$ on $\mathcal{X}^K$ with $\hat{\phi}(\mathbf{x}) \in \Phi$ for any sequence $\mathbf{x} = (x^1, x^2, \ldots, x^K)$ with $x^t \in \mathcal{X}$ for each t. The algorithm is described as follows: Choose an initial proposal distribution P and run a Metropolis-Hastings chain for K iterations with P as proposal distribution, That is, for each $t = 1, \ldots, K$ generate Y^t from P and accept Y^t, i.e. put $X^t = Y^t$, if $U^t \leq \min\{1, (\pi(Y^t)p(X^{t-1}))/(\pi(X^{t-1})p(Y^t))\}$. Otherwise, reject Y^t, i.e. put $X^t = X^{t-1}$. Here, p is the density of P, and the U^t's are independent samples from the uniform distribution on $[0, 1]$. Calculate $\phi_1 = \hat{\phi}(X^1, \ldots, X^K)$ and run according to the Metropolis-Hastings scheme another K iterations with P_{ϕ_1} as proposal. Calculate $\phi_2 = \hat{\phi}(X^{K+1}, \ldots, X^{2K})$ and repeat inductively. In addition, a convergence criterion is specified for the sequence $X^1, X^2, \ldots$. If satisfied for the first time at X^t, we put $M = [(t-1)/K]$, where $[\cdot]$ denotes integer value. This means that ϕ_M is the proposal distribution used to generate X^t (see Theorem 1). Alternatively, a convergence criterion may be specified directly in terms of the sequence $\phi_1, \phi_2, \ldots$. If satisfied for the first time at ϕ_m for some m, we put $M = m$ (see the last part of section 3). In any case, the proposal is then fixed at P_{ϕ_M} for the rest of the iterations, i.e. from $t = KM + 1$ onwards. This time point may then naturally be taken as the end of the burn in. The idea is to choose $\hat{\phi}$ in such a way that the closer $(x^1, \ldots x^K)$ is to represent a sample from π, the more should $P_{\hat{\phi}(x^1,\ldots,x^K)}$ resemble π. We will return to the question of what kind of "resemblance" that is desirable, and how this can be achieved. By fixing the proposal when the convergence criterion indicates that the last sequence is reasonably representative of π, we save computation time and obtain more stability and control of the estimation procedure.

Note that the transition probabilities at t in principle depend on all X^s for $s \leq K[t/K]$. This follows since at iteration no. t, the proposal distribution is $P_{\phi_{[t/K]}}$. Hence the stochastic process that has been defined, is not a Markov chain. Nevertheless, the process can be shown to converge under certain conditions:

Theorem 1 *Define $w_\phi(x) = \pi(x)/p_\phi(x)$, and suppose $w^*_\phi = \sup_x(w_\phi(x)) < \infty$ for all $\phi \in \Phi$, and also $w^* = \sup_\phi(w^*_\phi) < \infty$. Furthermore, define*

$$\tau = \min\{t : U^t \leq w_{\phi_{[(t-1)/K]}}(Y^t)/w^*_{\phi_{[(t-1)/K]}}\} \tag{1}$$

and let the proposal distribution be fixed at P_{ϕ_M} after τ, where $M = [(\tau - 1)/K]$. Then

(i) τ is stochastically dominated by a variable which is geometrically distributed with parameter $1/w^*$, and in particular, τ is finite with probability 1.

(ii) For any integers t, s with $t \leq s$ the distribution P^s of X^s satisfies $P^s(\cdot|\tau = t) = P^s(\cdot|\tau = t, X^0, \ldots, X^{t-1}) = \Pi(\cdot)$.

(iii) The distribution P^t of X^t satisfies $|P^t(A) - \Pi(A)| \leq (1 - 1/w^*)^t$ for any $A \subseteq \mathcal{X}$. In particular, P^t converges to π in total variation norm.

Proof: Suppose that at iteration no. t, p_ϕ is the proposal density, i.e. $\phi_{[(t-1)/K]} = \phi$. Clearly, $Pr(U^t \le w_\phi(Y^t)/w^*_\phi) = \int_{\mathcal{X}} (w_\phi(y)/w^*_\phi) p_\phi(y) dy = 1/w^*_\phi \ge 1/w^*$. This proves (i). Note that by definition, $(\tau = t)$ is equivalent to $(U^t \le w_\phi(Y^t)/w^*_\phi) \cap (\tau \ge t)$. Note also that $U^t \le w_\phi(Y^t)/w^*_\phi$ implies $X^t = Y^t$. Furthermore, $\phi_{[(t-1)/K]}$ is determined by $X_0, \ldots, X^{t-1}$, and given that $\phi_{[(t-1)/K]} = \phi$, the variables Y^t and U^t are independent of the variables $X^0, \ldots, X^{t-1}$ and the events $\tau \ne 1, \ldots, \tau \ne t-1$, whose intersection is the event $\tau \ge t$. Hence, we have

$$\begin{aligned}
P^t(A|\tau = t, X^0, \ldots, X^{t-1}) &= P^t(A|U^t \le w_\phi(Y^t)/w^*_\phi, \tau \ge t, X^0, \ldots, X^{t-1}) \\
&= P((X^t \in A) \cap (U^t \le w_\phi(Y^t)/w^*_\phi)|\tau \ge t, X^0, \ldots, X^{t-1}) / P(U^t \\
&\qquad \le w_\phi(Y^t)/w^*_\phi | \tau \ge t, X^0, \ldots, X^{t-1}) \\
&= P((Y^t \in A) \cap (U^t \le w_\phi(Y^t)/w^*_\phi)|\phi_{[(t-1)/K]} = \phi) / P(U^t \le w_\phi(Y^t)/w^*_\phi | \phi_{[(t-1)/K]} = \phi) \\
&= \Big(\int_A w_\phi(y) p_\phi(y) dy \Big) \Big/ \Big(\int_{\mathcal{X}} w_\phi(y) p_\phi(y) dy \Big) = \Pi(A) \; .
\end{aligned}$$

Now assume that (ii) is satisfied for some $s \ge t$. The candidate Y^{s+1} for X^{s+1} is generated by P_{ϕ_M}. For any ϕ, we denote by $P_\phi(x; \cdot)$ the transition kernel for a chain generated by the Metropolis-Hastings algorithm with P_ϕ as proposal distribution and Π as target distribution. Note that the random variable ϕ_M, given $\tau = t, X^0, \ldots, X^{t-1}$, is in fact a deterministic function of $X^1, \ldots, X^{t-1}$. Hence for any $A \subseteq \mathcal{X}$ we have $P^{s+1}(A|\tau = t, X^0, \ldots, X^{t-1}) = \int_{\mathcal{X}} p^s(x|\tau = t, X^0, \ldots, X^{t-1}) p_{\phi_M}(x; A) dx = \int_{\mathcal{X}} \pi(x) p_{\phi_M}(x; A) dx = \Pi(A)$ by the reversibility of the Metropolis-Hastings transition kernel $p_{\phi_M}(x; \cdot)$ with respect to π. Integrating with respect to $X^0, \ldots, X^{t-1}$ shows that $P^{s+1}(A|\tau = t) = \Pi(A)$. By induction, this proves (ii). By (i) and (ii), we have $|P^t(A) - \Pi(A)| \le P(\tau \le t)|P^t(A|\tau \le t) - \Pi(A)| + P(\tau > t)|P^t(A|\tau > t) - \Pi(A)| \le P(\tau > t) \le (1 - 1/w^*)^t$. Recall that the total variation distance is given by

$$|P^t - \Pi|_{TV} = \int_{\mathcal{X}} |p^t(x) - \pi(x)| dx = 2 \sup_{A \subseteq \mathcal{X}} |P^t(A) - \Pi(A)|. \tag{2}$$

Hence (iii) follows.

Note that part (ii) contradicts the assertion in Liu (1996) (cf. section 5) that the target distribution is never actually attained. Note also that by (ii), the AIC algorithm may be regarded as a procedure for exact sampling, with X^τ being an exact sample from π.

Suppose π is only known up to a proportionality constant, i.e. we know that $\pi = cf$ for some known function f and an unknown constant c. We may then define the ratios w_ϕ in terms of f instead of π and still detect the time τ for which the first exact sample X^τ is obtained. We still have a geometric rate of convergence, but the rate will equal $1/cw^*$ and will be unknown.

The boundedness condition on $w_\phi(x)$ will be discussed in appendix A. Of course, the convergence takes place even if one is not able to compute the suprema of $w_\phi(x)$, as long as this function is known to be uniformly bounded. In fact, the theorem remains valid if w^*_ϕ and w^* are replaced by any upper bounds c^*_ϕ and c^* with $c^*_\phi \le c^*$. This is important, since it may be difficult to compute w^*_ϕ and w^* explicitly. If we define $\tau_1, \tau_2, \ldots$ as the successive values of t for which $U^t \le w_{\phi_{[(t-1)/K)]}}(Y^t)/c^*_{\phi_{[(t-1)/K)]}}$, then $X^{\tau_1}, X^{\tau_2}, \ldots$ are independent samples from π. This observation will be very useful in the comparison with the rejection sampler in section 4. The times τ_i resemble the regeneration times used as times for adaption of the transition kernel in the scheme of Gilks, Roberts and Sahu (1998). But unlike the situation for the regeneration

times, the probability for the events $(\tau_i = t)$, $i = 1, 2 \ldots$ depend on the proposed value Y^t rather than on the state X^{t-1}. Another difference is that these events are determined by the same unitary U^t that decides acceptance or rejection of Y^t.

In the case of an IC, corresponding to Φ collapsing to a one point set $\{\phi\}$, and with $\hat{\phi}(x) = \phi$ for all $x \in \mathcal{X}$ the rate of convergence in Theorem 1 is the same as the rate given in Liu (1996). In the case of continuous distributions, Liu uses a coupling argument. It is worth noting that τ also may be considered as a coupling time, at least if there exists $x^* \in \mathcal{X}$ such that $w_\phi(x^*) = w_\phi^* = w^*$. Indeed, if one chain is started in x^*, while the initial value of another chain is sampled from π, and the two chains are linked by using the same proposals Y^t and the same uniform U^t to decide acceptance, then τ is the time the chains coalesce. Taking the Dirac measure at x^* as initial distribution, we also observe that $(1 - 1/w^*)^t$ is the smallest possible convergence rate for the total variation norm covering all initial distributions. To see this, note that if $X^0 = x^*$, then $\tau > t$ implies that $X^t = x^*$. Hence $P^t(\mathcal{X} - \{x^*\}|X^0 = x^*, \tau > t) = 0$. We then find by conditioning on the events $\{\tau > t\}$ and $\{\tau \leq t\}$ and using the continuity of π $\Pi(\mathcal{X} - \{x^*\}) - P^t(\mathcal{X} - \{x^*\}) = Pr(\tau > t) = (1 - 1/w^*)^t$. (See also Smith and Tierney (1996), section 4).

It is seen from Theorem 1 that in order to obtain fast convergence to π, one should look for P_ϕ for which $\pi(y)/p_\phi(y)$ stays small throughout all of $\mathcal{X}$. In particular, tail behaviour must be under control. After the chain has converged to π, acceptance rate and decay of autocovariances is more important, and may entail other requirements on P_ϕ. These characteristics of the algorithm, which are of interest in their own right, will be discussed in the rest of this section.

We consider a time $t \geq \tau$, so that X^t is π-distributed. For simplicity, we omit the index ϕ. Note that the inequality $w(y) \geq w(x)$ defines an ordering on $\mathcal{X}$. We put $A = \{(x, y) \in \mathcal{X}^2 : w(y) \geq w(x)\}, B = A^c$.

We first consider the acceptance rate a. Note that $a = 1$ is equivalent to $p = \pi$, so the Metropolis-Hastings algorithm would give independent samples from π. Hence, intuitively, a high acceptance rate is desirable. Also, practical experience indicates that a good estimate of the density function requires a certain number of different sample values from π, see Gåsemyr, Natvig and Sørensen (2000). This too makes a high acceptance rate favourable. We denote by A_x the set $\{y : (x, y) \in A\}$ and by B_y the set $\{x : (x, y) \in B\}$. We find the rejection rate

$$\begin{aligned}
1 - a &= 1 - \int_A \pi(x)p(y)dxdy - \int_B \pi(y)p(x)dx\,dy \\
&= \int_A \pi(x)\pi(y)dx\,dy + \int_B \pi(x)\pi(y)dx\,dy \\
&\quad - \int_{\mathcal{X}} \pi(x)(\int_{A_x} p(y)dy)dx - \int_{\mathcal{X}} \pi(y)(\int_{B_y} p(x)dx)dy \\
&\leq \int_{\mathcal{X}} \pi(x)(\int_{A_x} (\pi(y) - p(y))^+ dy)dx \\
&\quad + \int_{\mathcal{X}} \pi(y)(\int_{B_y} (\pi(x) - p(x))^+ dx)dy \leq |\pi - p|_{TV} \,.
\end{aligned}$$

Here, r^+ denotes the positive part of the real number r. Hence trying to minimize the total variation distance between proposal and target distributions would seem to result in a high acceptance rate.

We restrict our covariance analysis to one-step autocovariances. We assume without loss of generality that the function h for which we want to estimate $E_\pi(h(X))$, satisfies $E_\pi(h(X)) = 0$. We have

$$\begin{aligned}\mathrm{cov}(h(X^t), h(X^{t+1})) &= \int_B [h(x)^2(1 - w(y)/w(x)) \\ &\quad + h(x)h(y)w(y)/w(x)]\pi(x)p(y)dx\,dy + \int_A h(x)h(y)\pi(x)p(y)dxdy \\ &= \int_B h(x)^2\pi(x)p(y)dxdy + \int_B (-h(x)^2)\pi(y)p(x)dx\,dy \\ &\quad + \int_B h(x)h(y)\pi(y)p(x)dxdy + \int_A h(x)h(y)\pi(x)p(y)dx\,dy.\end{aligned}$$

Here $h(x)^2\pi(x)\pi(y)$ may be subtracted from the first integrand and added to the second. Since $E_\pi(h(X)) = 0$ implies that

$$\int_{\mathcal{X}^2} h(x)h(y)\pi(x)\pi(y)dxdy = 0\ ,$$

we may subtract $h(x)h(y)\pi(x)\pi(y)dxdy$ from the two last integrands to obtain

$$\begin{aligned}&\mathrm{cov}(h(X^t), h(X^{t+1})) = \\ &\quad = \int_B h(x)^2\pi(x)(p(y) - \pi(y))dxdy + \int_B h(x)^2\pi(y)(\pi(x) - p(x))dxdy \\ &\quad + \int_B h(x)h(y)\pi(y)(p(x) - \pi(x))dxdy + \int_A h(x)h(y)\pi(x)(p(y) - \pi(y))dxdy.\end{aligned}$$

If now $|h|$ is bounded by c, we obtain

$$\begin{aligned}&\mathrm{cov}(h(X^t), h(X^{t+1})) \le c^2 \int_{\mathcal{X}^2} \pi(x)|p(y) - \pi(y)|dxdy + c^2 \int_{\mathcal{X}^2} \pi(y)(\pi(x) - p(x))^+dxdy \\ &\quad + c^2 \int_{\mathcal{X}^2} \pi(y)|p(x) - \pi(x)|dxdy \le (5/2)c^2|\pi - p|_{TV}\ .\end{aligned}$$

If $|h|$ is decreasing (relative to the ordering defined by w), we have $|h(x)| \le| h(y)|$ for $(x, y) \in B$ and $|h(y)| \le| h(x)|$ for $(x, y) \in A$. In this case it therefore follows by a similar argument that $\mathrm{cov}(h(X^t), h(X^{t+1})) \le (5/2)\mathrm{var}_\pi(h(X))|\pi - p|_{TV}$. This monotonicity condition is of course very unlikely to be satified by coincidence, but if we have estimation of a particular h as a primary purpose of the analysis, we may try to search for proposal densities that satisfy this ordering condition. Thus, $p(x)$ should be large compared to $\pi(x)$ if $|h(x)|$ is large. But apart from this guideline for choosing p, it still seems to be useful to aim for a small total variation distance, especially if one is interested in estimating the expectations of several different functions.

3 Framework for constructing AIC algorithms, examples and heuristic arguments

We start this section by presenting a general framework for how the AIC algorithm may choose proposal distributions. Let $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_r)$ be an r-dimensional function on the set of distribu-

tion functions, whose components are characteristics of the distribution such as moments, quantiles or modulus. We will assume that the mapping $\phi \to \tilde{\theta}(P_\phi)$ is continuous. Put $\theta_0 = \tilde{\theta}(\Pi)$, and let $\Theta = \tilde{\theta}(\{P_\phi\}) \cup \{\theta_0\}$. The aim of the algorithm is to arrive at a P_{ϕ_M} which is close to Π, presumably achieved if P_{ϕ_M} have almost the same distribution characteristics as Π, i.e if ϕ_M satisfies

$$\tilde{\theta}(P_{\phi_M}) \approx \theta_0 \tag{3}$$

"Closeness" should preferably be with respect to both total variation distance and to the distance measure $d(P_\phi, \pi) = w^*_\phi$. To this end we construct a continuous function $\tilde{\phi} : \Theta \to \Phi$ for which we have:

(i) $\tilde{\phi}(\tilde{\theta}(P_\phi)) \approx \phi, \phi \in \Phi$ and also $\tilde{\theta}(P_{\tilde{\phi}(\theta)}) \approx \theta, \theta \in \Theta$.

(ii) By definition, $\tilde{\phi}(\theta_0) = \phi_0$.

(iii) The closer θ is to θ_0, the closer is $\tilde{\phi}(\theta)$ to ϕ_0 .

The function $\hat{\phi}$ introduced in the previous section is then defined by $\hat{\phi}(x^1, \ldots, x^K) = \tilde{\phi}(\hat{\theta}(x^1, \ldots, x^K)$, where $\hat{\theta}$ is an estimator for the relevant distribution characteristics. If for instance $\tilde{\theta}_i(P) = E_P(\psi_i(X))$ for some ψ_i, we may put $\hat{\theta}_i(x^1, \ldots, x^K) = (1/K)\sum_{l=1}^K \psi_i(x^l)$. Quantile estimates may also readily be obtained from the sample sequences of size K. In the special case of modulus estimates, a considerable extension of the framework may be useful. Indeed, the modulus could be estimated by monitoring the $\pi(Y^t)$, which must be known up to a proportionality constant and must be calculated anyway, for the proposed values $Y^t, t = 1, 2, \ldots$. The modulus estimate is changed to Y^t if $\pi(Y^t)$ is a new peak, exceeding all previous values.

Heuristically, we can argue that the adaptive algorithm should speed up convergence as follows: Let $Y^{Km+1}, \ldots, Y^{K(m+1)}$ be a sample from P_{ϕ_m}. Using (i), we should then have approximately $\hat{\phi}(Y^{Km+1}, \ldots, Y^{K(m+1)}) = \tilde{\phi}(\hat{\theta}(Y^{Km+1}, \ldots, Y^{K(m+1)})) \approx \tilde{\phi}(\tilde{\theta}(P_{\phi_m})) \approx \phi_m$. The Metropolis-Hastings acceptance-rejection procedure should ensure that $X^{Km+1}, \ldots, X^{K(m+1)}$ is more representative for π than $Y^{Km+1}, \ldots, Y^{K(m+1)}$ is, and hence by (iii) that $\phi_{m+1} = \hat{\phi}(X^{Km+1}, \ldots, X^{K(m+1)})$ is closer to ϕ_0 than $\phi_m \approx \hat{\phi}(Y^{Km+1}, \ldots, Y^{K(m+1)})$ is. More formally, we would have $|\phi_{m+1} - \phi_0|_d < |\phi_m - \phi_0|_d$. Here, $|\cdot|_d$ denotes some distance measure in R^d. Hence the algorithm should work its way towards the "optimal" proposal distribution P_{ϕ_0}, until it arrives at a ϕ_M close to ϕ_0 when the convergence criterion is satisfied. By the continuity of $\tilde{\theta}$, and by (ii) and (i), the distribution P_{ϕ_M} satisfies $\tilde{\theta}(P_{\phi_M}) \approx \tilde{\theta}(P_{\phi_0}) = \tilde{\theta}(P_{\tilde{\phi}(\theta_0)}) \approx \theta_0$, so that (3) is satisfied.

Remark 1 *It seems that existing proofs of convergence for adaptive algorithms that have been suggested, use some kind of uniform boundedness condition, satisfied either by assumption or by construction by the transition kernels or the samples they generate. The condition must be met for any possible history determining the transition kernels. This is the case for the convergence proofs in Gilks et al (1998), Hario et al (1998) and Holden (2000), as well as for our Theorem 1. In our view, it would be very attractive to construct convergence arguments actually taking advantage of the improvements in the ability of the transition kernels to produce approximate samples from* Π *which is intended by the adaption scheme. The above argument is an attempt in this direction. The most fundamental difficulty in formalising this argument*

is to be able to control the effect of transforming a sample of independent variables from P_ϕ by means of the Metropolis-Hastings accept-reject procedure.

The discussion so far in this section fits in with the parametric version (PAIC) presented in Gåsemyr, Natvig and Sørensen (2000). The non parametric version (NPAIC) can be covered by letting $\hat{\phi}$ be a density estimator and Φ the set of possible densities generated by $\hat{\phi}$. The heuristics can be adapted to cover this case as well.

Example 1. Suppose we may expect π to be reasonably well approximated by a gamma distribution with the correct expectation and variance. We may then define $\tilde{\theta}(P) = (E_P(X), \mathrm{var}_P(X)), \Theta = (0,\infty)^2$, and $p_\phi(x) = p_{a,b}(x) = g(x;a,b)$, where $g(x;a,b)$ denotes the gamma density with a, b respectively the shape and scale parameter. Hence, $\Phi = (0,\infty)^2$. With ξ and σ^2 representing respectively expectation and variance, we define $\tilde{\phi}(\xi,\sigma^2) = (\xi^2/\sigma^2, \xi/\sigma^2)$. In this case $\tilde{\phi}$ is a bijection between Φ and Θ, and $\tilde{\phi}(\tilde{\theta}(P_\phi)) = \phi$. We put $\phi_0 = \tilde{\phi}(\theta_0)$ by definition, and we then have that P_{ϕ_0} and π have the same expectation and variance. The monotonicity of distances (iii) can be achieved by choosing some metric, e.g. the Eucledian metric, for Φ, and defining the distance of θ and θ' in Θ as the corresponding distance in Φ between $\tilde{\phi}(\theta)$ and $\tilde{\phi}(\theta')$. With these definitions, this example fits very well into the general set up. Different versions of this example have been studied in Gåsemyr, Natvig and Sørensen (2000), with π representing the posterior distribution for the failure rates in exponential survival models with data containing left censorings. Extensions of the example where $n > 1$ and P_ϕ is a product of n gamma distributions, are also studied. The results are very promising. With execution times comparable to any of the other simulation techniques that were tried, samples were produced that could be used to produce very good approximations to the marginal densities of π.

If in some model another two-parametric class of distributions is expected to approximate π better than the class of gamma distributions, one may clearly copy the procedure described in example 1 with this class instead of the gamma distributions.

In the examples studied in Gåsemyr, Natvig and Sørensen (2000), the condition $w^* < \infty$ is not satisfied, so theoretically convergence is not assured. Nevertheless, the AIC algorithm worked very well in practice. It was also observed that the scale and shape parameters of the proposal distributions displayed a nice, monotonic convergence to the target values. This suggests that the heuristic argument above may be a better description of the mechanism behind the algorithm than the theoretical proof of Theorem 1.

However, the boundedness condition may in many cases be obtained by allowing ϕ to run in only a subset Φ_1 of the possible parameter values. Hence, assume that $w_1^* = \sup_{\phi\in\Phi_1}(w_\phi^*) < \infty$. Correspondingly define a function $\tilde{\phi}_1 : \Theta \to \Phi_1$ still satisfying (i), (ii) and (iii). Replacing $\tilde{\phi}$ by $\tilde{\phi}_1$ in the construction of an AIC gives an algorithm for which Theorem 1 is valid. Restricting the parameter space in this way, one must expect less accuracy in the approximations (i). To compensate for this, one may as a compromise choose a mixture distribution $\alpha P_{\phi_1} + (1-\alpha)P_{\phi_2}$, with parameter space correspondingly extended to $\Phi = \Phi_1 \times \Phi_2$ and a function $\tilde{\phi} = (\tilde{\phi}_1, \tilde{\phi}_2)$. Here $\Phi_1, \tilde{\phi}_1$ are as above, Φ_2 is the full original parameter space, $\tilde{\phi}_2$ chooses a best possible match within $\{P_\phi, \phi \in \Phi_2\}$ for the distribution characteristics $\theta_1, \ldots, \theta_r$ and α is a fixed weight, $0 < \alpha \leq 1$. More generally, α may be allowed to depend on the sample, i.e. $\alpha = \tilde{\alpha}(\hat{\theta}(x^1, \ldots, x^K))$, but must at least be bounded below by a fixed $\beta > 0$ in order to ensure $w^* < \infty$. Indeed, we obtain $w^* \leq (1/\beta)w_1^* < \infty$. Accordingly, the parameter space may be taken as $\Phi = \Phi_1 \times \Phi_2 \times [\beta, 1]$.

Suppose in particular that $\pi(x) \propto \pi_0(x)L_x(D)$, where π_0 is a prior distribution and $L_x(D)$ is the likelihood for x obtained from data D. Then the boundedness condition $w^* < \infty$ is ensured if p_{ϕ_1} is replaced by π_0 in the above set up and $L_x(D)$ is bounded. In Bayesian analysis the prior distribution is often easy to simulate from, so that the conditions ensuring convergence in Theorem 1 can often be met in a practical way.

The burn in of the AIC algorithm may be taken as the iterations performed before X^t can be considered as either exactly or approximately π-distributed. It is natural to fix the proposal distribution after the burn in. For if $X^{Km+1}, \ldots, X^{K(m+1)}$ are approximately π-distributed, then ϕ_m is almost as good an approximation to ϕ_0 as ϕ_k would be for any $k > m$. Another reason to fix the proposal after some time is that as ϕ_m approaches ϕ_0, the expected size of the moves towards ϕ_0 of the sequence $\{\phi_m\}$ must necessarily drop, and therefore get outpowered by the Monte Carlo variance at some stage. Theorem 1 indicates a burn in of length $K[(\tau - 1)/K]$. However, if $w^* = \infty$ this does not work. In addition, Theorem 1 may suggest a far too long burn in if the real convergence rate is governed by the mechanism described in our heuristic convergence argument. Instead we suggest as a diagnostic test to stop the burn in at $t = K(m + 1)$ if $|\phi_m - \phi_{m+1}|_d$ is sufficiently small. Hence, we use P_{ϕ_M} with

$$M = 1 + \inf\{m : |\phi_m - \phi_{m+1}|_d \leq \delta\} \tag{4}$$

as proposal for the rest of the iterations. For instance, we may put $|\phi_m - \phi_{m+1}|_d = \sup_{1 \leq i \leq d}\{|\phi_{m,i} - \phi_{(m+1),i}|/\epsilon_i\}$, where ϵ_i are scaling constants. Using the criterion $|\phi_m - \phi_{m+1}|_d \leq 1$, the result is that the burn in is finished when for the first time $|\phi_{m,i} - \phi_{(m+1),i}| \leq \epsilon_i$ for all $i = 1, \ldots, d$. Such a criterion is used successfully in the examples in Gåsemyr, Natvig and Sørensen (2000). In less nicely behaved applications, the criterion might be met prematurely by chance, resulting in a too short burn in. This could be remedied by requiring the criterion to be met for several consecutive values of m.

The procedure for determining the burn in described above focuses on the distance between components of the parameter vector ϕ for the consecutive proposal distributions. An alternative is to focus on the distance of components of the vector θ of distribution characteristics. This would seem particularly sensible if θ consists of moments and a primary purpose of the analysis is to estimate moments of π. The result would be that when the distances between moment estimates, based on consecutive sequences, are small enough, the burn in is terminated. Such a criterion for ending the burn in does not ensure that the chain has converged to the stationary distribution π, but it would nevertheless seem reasonable to believe that subsequent samples represent features of π relevant for moment estimation sufficiently well. These considerations are relevant also if the aim is to estimate $E_\pi(h(X))$ if h is well approximated by the first few terms of the Taylor series expansion within most of the support of π.

Numerous modifications of the framework outlined in this section could be made. An obvious and probably sensible possibility is to use parameter values of the form $(1/l)(\phi_{m+1} + \cdots + \phi_{m+l})$ for some $l > 1$, both as a basis for the diagnostic test and as the parameter for the final proposal distribution. Presumably, this would diminish the sensitivity of the algorithm to Monte Carlo variability. The same effect could be achieved by replacing the fixed sequence length K by an increasing $K(m)$. In this way the dominance of Monte Carlo variation over the size of the expected move towards ϕ_0 as ϕ_m approaches ϕ_0 may be defeated. This is the basis for a proof of convergence that does not require the boundedness condition, see Theorem 2 in appendix B. The function $K(m)$ may be deterministic or random, depending on the history. In practice,

the choice of such a function would have to be done by trial and error, which would probably not be worthwhile if the procedure is constructed only for the analysis of a particular problem. If on the other hand the procedure is meant for regular use with varying input, investment in the search for a suitable function $K(m)$ may be justified.

4 Comparison of rejection sampling, independent chains and importance sampling

In this section we consider a fixed proposal distribution, and derive a bound on the asymptotic efficiency of an independent chain relative to rejection sampling using the same candidate distribution. This is the relevant criterion for asymptotic comparison also of the AIC with rejection sampling, since the proposal distribution of the AIC is fixed after the burn in.

Define $\mu = E_\pi(h(X))$ and $\sigma^2 = \mathrm{var}_\pi(h(X))$. Let P be a fixed proposal distribution with density p, and suppose that $w^* = \sup_{\mathcal{X}}(w(x))$ is finite and known. Then a sequence of N independent pairs (Y^t, U^t), where Y^t is P-distributed, and U^t is uniform on $[0,1]$ and independent of Y^t, can be used as a basis for estimates of μ both using rejection sampling and an independent chain Metropolis-Hastings algorithm. It is convenient to consider N as random, equal to the smallest number τ_m of samples needed to obtain m independent draws from π using the rejection sampler, where m is any integer > 0. We denote the corresponding sample mean estimate of μ by $\hat{\mu}^R(m)$. Based on the Markov chain X^t obtained from (Y^t, U^t), we can define the sample mean estimate for μ as $\hat{\mu}^M(m) = (1/\tau_m)\sum_{t=1}^{\tau_m} h(X^t)$. We want to compare these two estimates asymptotically, and we may assume that the initial state X^0 of the Markov chain is π-distributed. A natural criterion for comparison of rejection sampling with the independent chain is then the size of $\rho = \lim_{m\to\infty}\rho(m)$, where $\rho(m) = \mathrm{var}(\hat{\mu}^M(m))/\mathrm{var}(\hat{\mu}^R(m))$.

Now put $M(0) = 0$ and $M(t) = \sum_{s=1}^{t} I(U^s \le w(Y^s)/w^*) =$ the number of independent drawings from π obtained after t samples from P for $t = 1, 2, \ldots$. Also put $\tau_0 = 0, \tau_i = \min\{t : M(t) = i\}, i = 1, \ldots, m$. Also, define $R_i = \tau_i - \tau_{i-1}, i = 1, \ldots, m$. The variables $X^1, \ldots, X^{\tau_m}$ may be grouped into independent segments $\{X^1, \ldots, X^{\tau_1 - 1}\}, \{X^{\tau_1}, \ldots, X^{\tau_2 - 1}\}, \ldots, \{X^{\tau_m}\}$ with respectively $R_1 - 1, R_2, R_3, \ldots, R_m, 1$ variables. The variables $R_i, i = 1, \ldots, m$ are independent and geometrically distributed with parameter $1/w^*$. We have

$$N = \tau_m = \sum_{i=1}^{m} R_i \tag{5}$$

Define $\hat{\mu}_i = (1/R_i)\sum_{t=\tau_{i-1}}^{\tau_i - 1} h(X^t), i = 2, 3, \ldots, m$. For $i = 1$ the expression is for convenience slightly modified by replacing $h(X^0)$ by $h(X^{\tau_m})$. This gives

$$\hat{\mu}^M(m) = (1/\tau_m)\sum_{i=1}^{m} R_i\hat{\mu}_i \tag{6}$$

Clearly, the $\hat{\mu}_i$'s are independent. On the other hand, we make no assumptions on the covariance structure within each segment $\{X^{\tau_i}, X^{\tau_i+1}, \ldots, X^{\tau_{i+1}-1}\}$ of the Markov chain. This means that we may have $\mathrm{cov}(h(X^t), h(X^{t+s})) = \sigma^2$ given that $\tau_i \le t < t + s < \tau_{i+1}$ for some i, indicating

a deterministic dependence between samples from the same segment. Hence, we only base our comparison on the very conservative bound

$$\mathrm{var}(\hat{\mu}_i) \leq \sigma^2 \tag{7}$$

Using (6), (7) and (5), the independence of the $\hat{\mu}_i$'s and the symmetry of the $R_i, i = 1, \ldots, m$ this gives

$$\begin{aligned}
\rho(m) &= m\mathrm{var}(\hat{\mu}^M(m))/\sigma^2 = (m/\sigma^2)[E(\mathrm{var}(\hat{\mu}^M(m)|\tau_1, \ldots, \tau_m)) + \mathrm{var}(E(\hat{\mu}^M(m)|\tau_1, \ldots, \tau_m))] \\
&= (m/\sigma^2)[E((1/\tau_m^2)(\sum_{i=1}^{m} R_i^2 \mathrm{var}(\hat{\mu}_i|\tau_1, \ldots, \tau_m)) \\
&\quad + \mathrm{var}(\mu|\tau_1, \ldots, \tau_m)] \leq mE((1/\tau_m^2)(\sum_{i=1}^{m} R_i^2)) = m^2 E((1/\tau_m^2)R_1^2) \leq m^2 E((R_1^2/(\sum_{i=2}^{m} R_i)^2) \\
&= E(R_1^2)E(((1/m)\sum_{i=2}^{m} R_i)^{-2}).
\end{aligned}$$

Since R_1 is geometrically distributed with parameter w^*, the first factor is $2(w^*)^2 - w^*$, while the second factor tends to $(1/w^*)^2$ by the strong law of large numbers and the bounded convergence theorem. Hence, we get

$$\rho \leq 2 - 1/w^* \tag{8}$$

In fact, this matches exactly the result obtained in Liu (1996) in the case of finite state spaces, and shows that rejection sampling may potentially be twice as efficient as independent Metropolis sampling. But remember that our result is based on assuming $\mathrm{cov}(h(X^t), h(X^{t+s})) = \sigma^2$ given that $\tau_i \leq t < t + s < \tau_{i+1}$ for some i. By a reasonable decay of autocovariances, the independent Metropolis sampling will be much more efficient. Furthermore, if w^* has to be replaced by an upper bound c^*, the efficiency of independent Metropolis-Hastings sampling remains unchanged, whereas the efficiency of the rejection method will be reduced. Moreover, the Metropolis-Hastings sampling will be even further improved by allowing for adaptivity if this leads to a reduction in autocovariances and in w_ϕ^* for the final proposal density p_ϕ.

If we modify the estimate for μ based on the IC to $\hat{\mu}_1^M(m) = (1/m)\sum_{i=1}^{m} \hat{\mu}_i$, we obtain a corresponding ratio of variances $\rho_1(m)$ satisfying $\rho_1(m) \leq 1$. The fact that we may have $\rho(m) > 1$ for the standard estimate $\hat{\mu}^M(m)$ is accounted for by the extra variability due to the random weights $R_i/(\sum_{j=1}^{m} R_j)$ allotted to the $\hat{\mu}_i$. This does not necessarily mean that the estimate $\hat{\mu}_1^M(m)$ using fixed weights $1/m$ is better in practice.

Remark 2 *Relation to importance sampling.*

The variable Y^t sampled from p at time t is represented in the resulting IC a random number $W^t(Y^t)$ times. We will show that if the chain is started at stationarity, then

$$E(W^t(Y^t)|Y^t = y) = w(y) \ ,$$

so that Y^t is represented by a weight whose expected value is the same as the weight that would have been used in importance sampling using the same proposal distribution. To see this, denote

by P and Π the cumulative distribution functions associated with p and π respectively, where the ordering of $\mathcal{X}$ is determined by the value of $w(y), y \in \mathcal{X}$. The probability of accepting y when proposed is $q(y) = \int_{x<y} \pi(x)dx + \int_{x \geq y}(w(y)/w(x))\pi(x)dx = \Pi(y) + w(y)\bar{P}(y)$. On the other hand, if the current state is y, the probability of rejecting the next proposal is $\lambda(y) = \int_{y'<y}(1-w(y')/w(y))p(y')dy' = P(y) - (1/w(y))\Pi(y)$. The distribution of $W^t(Y^t)$ given $Y^t = y$ equals the distribution of ZS, where Z, S are independent, Z is Bernoulli with parameter $q(y)$ and S is geometric with parameter $1 - \lambda(y) = \bar{P}(y) + (1/w(y))\Pi(y) = (1/w(y))q(y)$. This gives $E(W^t(Y^t)|Y^t = y) = E(Z)E(S) = w(y)$ as asserted.

In principle, one could make a comparison of the asymptotic efficiency of the IC and importance sampling based on the framework given in this proof. This would involve studying the covariance structure of the IC by means of the pairs $(Y^t, W^t(Y^t))$ rather than directly on the X^t-chain. We have not been able to obtain any striking results by persuing this idea, however.

5 High-dimensional models

Even though the heuristic arguments of section 3 indicate that the proposal distribution will approach the "optimal" P_{ϕ_0} when t increases, the performance of the AIC algorithm depends crucially on the possibility of approximating the target distribution sufficiently well by such a parametric distribution P_{ϕ_0}. The higher the dimension of $\mathcal{X}$ is, the more difficult this would seem to be. This problem is studied in this section. We briefly consider heuristically the expected behaviour of the AIC as dimension increases, and present some ideas concerning the construction of proposal distributions. A modified version of the AIC, called CAIC, (componentwise AIC) emerges from this general discussion. We also present an example showing how one could go about constructing proposal distributions in practice in a model of moderately high dimension.

As stated earlier, there is both heuristic and experimental evidence that total variation distance between target and proposal distributions is the most important feature of the approximation. Recall that $|P_1 - P_2|_{TV}$ denotes the total variation distance between the distributions P_1, P_2 (cf. (2)). For any distribution $P(x, y)$ we denote by P_X and $P_{Y|x}$ the marginal distribution of X and the conditional distribution of Y given $X = x$ respectively. Now if $n = 2$ and $\mathcal{X} = R^2$ we have for any proposal distribution P with density p that

$$\begin{aligned} |P - \Pi|_{TV} &= \int_{R^2} |\pi(x_1, x_2) - p(x_1, x_2)|dx_1 dx_2 \leq \int_{R^2} \pi_{X_2|X_1}(x_2|x_1)|\pi_{X_1}(x_1) - p_{X_1}(x_1)| \\ &+ \int_{R^2} p_{X_1}(x_1)|\pi_{X_2|X_1}(x_2|x_1) - p_{X_2|X_1}(x_2|x_1)|dx_1 dx_2 \\ &= |P_{X_1} - \Pi_{X_1}|_{TV} + \int_R p_{X_1}(x_1)|P_{X_2|x_1} - \Pi_{X_2|x_1}|_{TV} dx_1 \ . \end{aligned}$$

By replacing X_1 by $(X_1, \ldots, X_{n-1})$ and X_2 by X_n it follows by induction that we have for arbitrary n that

$$\begin{aligned} |P - \Pi|_{TV} \leq |P_{X_1} - \Pi_{X_1}|_{TV} + \cdots + \int_{R^{n-1}} p_{X_1,\ldots,X_{n-1}}(x_1, \ldots, x_{n-1})|P_{X_n|x_1,\ldots,x_{n-1}} \\ -\Pi_{X_n|x_1,\ldots,x_{n-1}}|_{TV} dx_1 \cdots dx_{n-1} \end{aligned} \tag{9}$$

The significance of (9) is perhaps most evident when considering a model class where new dimensions can be added within a common symmetric structure. The inequality (9) then expresses a roughly linearly increasing bound on the total variation distance as the dimensionality of the model increases. This does not indicate a more serious problem with high dimensionality for the AIC algorithm in general than for e.g. the Gibbs sampler or versions of the Metropolis-Hastings algorithm treating one component of x at a time.

The sequential conditioning structure in (9) also suggests a way of choosing proposal distributions in many problems, in particular in models involving discrete time. It will then often be natural to choose a proposal distribution of the form $P_\phi(x) = P_{\phi_1}(x_1)P_{\phi_2}(x_2|x_1)\cdots P_{\phi_n}(x_n|x_1, \ldots, x_{n-1})$ In this situation the parameter vector $\phi = (\phi_1, \ldots, \phi_n)$ may be selected componentwise by choosing distribution characteristics $\tilde{\theta}^k = (\tilde{\theta}^k_1, \ldots, \tilde{\theta}^k_r), k = 1, \ldots, n$ such that $\tilde{\theta}^k(P)$ primarily reflects the distributional aspects of X_k given $X_1, \ldots, X_{k-1}$ under P. We define corresponding estimators $\hat{\theta}^k, k = 1, \ldots, n$ based on samples of size K, and obtain proposal distribution parameters of the form $\hat{\phi}(x^1, \ldots, x^K) = (\tilde{\phi}_1(\hat{\theta^1}(x^1, \ldots, x^K)), \ldots, \tilde{\phi}_n(\hat{\theta^n}(x^1, \ldots, x^K)))$ for suitably chosen functions $\tilde{\phi}_k, k = 1, \ldots, n$.

As a simple illustration, suppose that the number of cases of a decease, occurring in a particular population in time intervals $(t_{k-1}, t_k), k = 1, \ldots, n$ of equal length, are Poisson distributed with unknown parameters λ_k. Let $x_k = \lambda_k$, and let Π be the posterior distribution of $x = (x_1, \ldots, x_n)$ given data D. Even in such a seemingly simple model, the data D may be in a form that gives a very complicated likelihood function, and the computational challenge may be formidable, see e.g. Glad et al. (2000). One possibility for proposal distribution is to let P_{ϕ_k} be a normal distribution with a trend adjusted expectation and a variance common for all k. Hence, $E_{P_{\phi_k}}(X_k|x_1, \ldots, x_{k-1}) = x_{k-1} + \phi_k$. In this situation, it is natural to choose $\tilde{\theta^k}(P) = E_P(X_k - X_{k-1})$, which is readily estimated by the corresponding sample mean, and $\tilde{\phi_k}(\theta^k) = \theta^k$.

From the sequential procedure for selecting components of ϕ described above, it is a short step to an adaptive, non-Markovian version of a more traditional Metropolis-Hastings algorithm processing one component of x at a time. To construct such an algorithm, we should like to choose a P_{ϕ_k} that approximates $\Pi_{X_k|X_l, l\neq k}$. The distribution characteristics $\tilde{\theta^k}$ should then reflect the conditional distribution of X_k given all other components of x, not only $X_1, \ldots, X_{k-1}$. Each proposed x_k drawn from P_{ϕ_k} is exposed to the usual Metropolis-Hastings accept-reject step separately. We still base the selection of parameters ϕ_k on estimates based on sequences of length K of previous iterations, however, so that the Markov property is not satisfied. In the Poisson model described above, one could e.g. obtain such an algorithm by means of the same basic parameters, only replacing x_{k-1} by $(x_{k-1} + x_{k+1})/2$ in the definitions.

The algorithm just described differs from an AIC by processing one component at a time. A more essential difference is that the proposal distributions do not stay fixed throughout the sequence of K iterations, since the proposal distribution for the k-th component depends on $x_l, l \neq k$. On the other hand, it resembles the AIC by not depending on the value of x_k from the previous iteration. In these respects, the algorithm also resembles the Gibbs sampler, and may be viewed as an adaptive approximation to the Gibbs sampler in situations where it is impossible or at least very difficult to sample directly from the conditional distributions, as required in the Gibbs sampler. This approximation is closer the better $P_{\phi_k}(x_k|x_l, l \neq k)$ approximates $\Pi(x_k|x_l, l \neq k)$. We will term the algorithm a componentwise adaptive independent chain

(CAIC).

At present we can not give any general recipy for choosing proposal distributions in the AIC in any particular situation. The Poisson example described above may give some ideas. We recall from the example in section 3 that another possibility is to estimate the marginal expectation and variance for each component of x, derive proposal distributions for each component within a certain two-parameter parametric class with the corresponding moments, and use the product of these distributions as proposal. It should also be mentioned that if a multivariate normal distribution seems like a suitable choice for proposal distribution, then the expectation and the covariance matrix may be determined by means of estimates based on previous samples. However, measures must be taken to ensure a positive definite covariance matrix. Section 4.2 of Gilks et al. (1998) also presents examples where a Gaussian proposal distribution is updated in this way, based on their "adaption at regeneration times" scheme with independent Gaussian proposals. In Haario et al. (2000) a similar idea is used, but their algorithm is based on random walk proposals, and also differs by using the entire history in the estimation of covariance matrices (versions of the algorithm using only the last K iterations are also presented, but do not necessarily converge). To give further ideas for the selection of proposal distributions, we conclude this section by a somewhat more complicated example.

Example 2. This example is taken from Arjas an Gasbarra (1996), and we have chosen to stick essentially to their notation. This is partly in conflict with the notation used elsewhere in this paper. The aim is to obtain samples from $\Pi(\lambda)$, the posterior distribution for $\lambda = (\lambda_1, \ldots, \lambda_n)$, a parameter vector representing a piecewise constant approximation to the hazard rate $\lambda(t), t_0 \leq t \leq t_n$ for failure of some sort, e.g. the occurence of a decease, or death due to it, among individuals in a certain population. Here, λ_k is the hazard rate on the interval $(t_{k-1}, t_k]$, where $t_0 < t_1 < \ldots < t_n$ and $t_k - t_{k-1} = (1/n)(t_n - t_0)$. The data D are of the form $(X_i, \delta_i), i = 1, \ldots, N$, obtained from a study population of N individuals. Here $X_i =$ the time at which the i-th individual was last seen and δ_i is the indicator function for the i-th individual failing at X_i. The prior distribution is defined as follows: Let $g(\cdot; a, b)$ be the gamma distribution with shape parameter a and scale parameter b. Fix parameters $\alpha_0, \beta_0, \alpha$. Let λ_1 be distributed according to $g(\cdot; \alpha_0, \beta_0)$. Given $\lambda_1, \ldots, \lambda_{k-1}$, let λ_k be distributed according to $g(\cdot; \alpha, \beta_k)$, where $\beta_k = \alpha/\lambda_{k-1}$. This means that $E_{\Pi_0}(\lambda_k|\lambda_1, \ldots, \lambda_{k-1}) = \lambda_{k-1}$. Defining $Y(t) =$ the number of individuals at risk at time t and $R_k = \sum_{i=1}^{N} I(t_{k-1} < X_i \leq t_k)\delta_i$, the likelihood is given by $L(\lambda|D) = \prod_{k=1}^{n}[\lambda_k^{R_k} \exp(-\lambda_k \int_{t_{k-1}}^{t_k} Y(s)ds)]$.

Our suggestion for proposal distribution is to use the sequential conditioning framework described in this section with each $P_{\phi_k}(\cdot|\lambda_1, \ldots, \lambda_{k-1})$ being a gamma distribution with shape and scale parameters respectively being α_k and $\gamma_k(\lambda_{k-1}) = \beta_k + \eta_k = \alpha/\lambda_{k-1} + \eta_k$. In order to make this definition cover the case $k = 1$, we define $\lambda_0 = \alpha/\beta_0$, implying that $\beta_1 = \beta_0$. The procedure for selecting $\phi_k = (\alpha_k, \eta_k)$ is motivated by a wish to obtain approximate equality of conditional first and second moments, i.e.

$$E_\Pi(\lambda_k|\lambda_{k-1}) \approx E_{P_{\phi_k}}(\lambda_k|\lambda_{k-1}) = \alpha_k/(\alpha/\lambda_{k-1} + \eta_k) \qquad \text{and}$$

$$E_\Pi(\lambda_k^2|\lambda_{k-1}) \approx E_{P_{\phi_k}}(\lambda_k^2|\lambda_{k-1}) = (\alpha_k(\alpha_k + 1))/(\alpha/\lambda_{k-1} + \eta_k)^2 \, .$$

To derive equations for determination of α_k and η_k from these approximate equalities, we multiply with the denominators on the right hand side and take expectations with respect to

λ_{k-1} to obtain

$$\alpha_k \approx \alpha E_\Pi(\lambda_k/\lambda_{k-1}) + \eta_k E_\Pi(\lambda_k) \tag{10}$$

and

$$\alpha_k(\alpha_k+1) \approx \alpha^2 E_\Pi((\lambda_k/\lambda_{k-1})^2) + 2\alpha\eta_k E_\Pi(\lambda_k^2/\lambda_{k-1}) + \eta_k^2 E_\Pi(\lambda_k^2) \tag{11}$$

We now replace the distribution characteristics $\tilde{\theta_i^k}(\Pi), i = 1, \ldots, 5$, being respectively $E_\Pi(\lambda_k)$, $E_\Pi(\lambda_k/\lambda_{k-1})$, $E_\Pi(\lambda_k^2)$, $E_\Pi(\lambda_k^2/\lambda_{k-1})$ and $E_\Pi((\lambda_k/\lambda_{k-1})^2)$ by corresponding estimates $\hat{\theta_i^k}$ and the $\approx$ signs by equality signs to obtain a pair of equations that can be solved by inserting the right hand side of the first equation for α_k in the second.

It is worth noting that following this procedure, it is, at least theoretically, possible to obtain exact samples by restricting the parameter space Φ. Indeed, some standard calculations show that we have

$$w_\phi(\lambda) \propto \Big[\prod_{k=2}^{n-1} \lambda_k^{R_k-\alpha_k+\alpha_{k+1}}(\alpha+\lambda_k\eta_{k+1})^{-\alpha_{k+1}} e^{-\lambda_k(\int_{t_{k-1}}^{t_k} Y(s)ds-\eta_k)}\Big]$$
$$\times \lambda_1^{R_1-\alpha_1+\alpha_2+\alpha_0-\alpha}(\alpha+\lambda_1\eta_2)^{-\alpha_2} e^{-\lambda_1(\int_{t_0}^{t_1} Y(s)ds-\eta_1)} \lambda_n^{R_n+\alpha-\alpha_n} e^{-\lambda_n(\int_{t_{n-1}}^{t_n} Y(s)ds-\eta_n)}$$

Note that this product form arises due to cancellations of equal terms from $\pi(\lambda)$ and $p_\phi(\lambda)$, despite the dependence of $\lambda_1, \ldots, \lambda_n$ under both these densities. By suitably restricting the parameter space Φ, this can be maximized term by term with respect to λ. We must necessarily have the restriction $\eta_k \le \int_{t_{k-1}}^{t_k} Y(s)ds$, $k = 1, \ldots, n$. The restriction for $(\alpha_1, \ldots, \alpha_n)$ is somewhat more arbitrary, but it is very natural to impose $\alpha \le \alpha_k \le \alpha + R_k$. With such a restricted Φ, Theorem 1 applies and permits exact sampling. One may also use a mixture distribution with one component based on a restricted parameter space and one based on the unrestricted $(R^+)^{2n}$.

Note that these upper bounds for η_k and α_k correspond to updating of scale and shape parameters of the prior distribution for λ_k, given $\lambda_1, \ldots, \lambda_{k-1}$, with respectively the total time on test and the number of observed failures that correspond to the interval $(t_{k-1}, t_k]$ and is summarised in that part of the likelihood that involves λ_k. This would be the correct way of updating the parameters had $\lambda_1, \ldots, \lambda_n$ been independent a priori. Using this simple update may be a good choice for an initial proposal distribution.

Arjas and Gasbarra themselves use a Metropolis-Hastings algorithm that is in some respects similar to the CAIC algorithm introduced above. Like the CAIC, their algorithm processes one component of λ at a time, and the proposal distribution for λ_k is allowed to depend on $\lambda_i, i \neq k$, but not on the old value of λ_k. On the other hand, the proposal distribution does not adapt to what could be learned from previous iterations. The proposal is a gamma distribution whose modulus m_k is the same as for $\pi(\lambda_k|\lambda_i, i \neq k)$. A CAIC could be based on the same gamma proposals as above, i.e. with parameters derived from equations based on (10) and (11). Alternatively, one could replace (10) by $(\alpha_k - 1)/(\alpha/\lambda_{k-1} + \eta_k) = m_k$, so that the proposal has the same modulus as in the Arjas - Gasbarra algorithm.

We emphasize that the algorithm used by Arjas and Gasbarra (1996) seems to work quite well and is not in need of replacement. In fact, the model described here is only a building block in the cumulative hazard rate ordering problem they are actually considering. The purpose of our alternative suggestions is to present ideas that may have more general applicability.

Acknowledgement

We are very greatful to Professor Bent Natvig for his careful reading of the manuscript. The presentation of the material has improved substantially thanks to his valuable suggestions.

References

Arjas, E., and Gasbarra, D., "Bayesian inference of survival probabilities under stochastic ordering constraints," J. Amer. Statist. Ass. 91 (435), 1101–1109, 1996.

Gilks, W.R., Roberts, G.O. and Sahu, S.K., "Adaptive Markov Chain Monte Carlo through regeneration," J. Amer. Statist. Ass. 93, 1045–1054, 1998.

Gåsemyr, J., Natvig, B. and Sørensen, E., "A comparison of two sequential Metropolis-Hastings algorithms with standard simulation techniques in Bayesian inference in reliability models involving a generalized gamma distribution," To appear in Methodology and Computing in Applied Probability.

Glad, I.K., Frigessi, A., Scala Tomba, G., Balducci, M. and Pezzotti, P., "Bayesian back calculation with HIV seropositivity notifications," To appear in Biometrics.

Haario, H., Saksman, E. and Tamminen, J., "An adaptive Metropolis algorithm," Department of Mathematics, P. O. Box 4, FIN–00014, University of Helsinki, Finland, 2000.

Hastings, W. K., "Monte Carlo sampling methods using Markov chains and their applications," Biometrica 57, 97–109, 1970.

Holden, L., "Adaptive chains," Submitted J. Roy. Stat. Soc. Ser. B.

Liu, J. S., "Metropolized independent sampling with comparison to rejection sampling and importance sampling," Statistics and Computing 6, 113–119, 1996.

Propp, J. G. and Wilson, D. B., "Exact sampling with coupled Markov chains and applications to statistical mechanics," Random structures and algorithms 9, 223–252, 1996.

Smith, R.L. and Tierney, L., "Exact transition probabilities for the independence Metropolis sampler," Technical report, Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27599–3260, USA, 1996.

Tierney, L., "Markov chains for exploring posterior distributions," Ann. Statistics 22, 1701–1762, 1994.

6 Appendix A Sufficient conditions for Theorem 1 and counterexample.

The following proposition gives a sufficient condition for the boundedness of $w_\phi(x)$, ensuring the validity of Theorem 1.

Proposition 1 *Suppose that Φ is compact, and that for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $x \in \mathcal{X}$, we have $|w_\phi(x) - w_{\phi'}(x)| < \epsilon$ whenever $|\phi - \phi'| < \delta$. Then $w_\phi(x)$ is bounded on $\Phi \times \mathcal{X}$.*

Proof: We show that w_ϕ^* is continuous in ϕ. The assertion then follows from the compactness of Φ.

Let $\phi \in \Phi$ be arbitrary. Choose $x \in \mathcal{X}$ such that $w_\phi(x) > w_\phi^* - \epsilon/2$. By assumption, there exists $\delta > 0$, which may be chosen independently of ϕ, such that if $|\phi' - \phi| < \delta$, then $|w_{\phi'}(x) - w_\phi(x)| < \epsilon/2$. Hence $|\phi' - \phi| < \delta$ implies that $w_{\phi'}^* > w_\phi^* - \epsilon$. By symmetry we also have $w_\phi^* > w_{\phi'}^* - \epsilon$. Hence w_ϕ^* is in fact uniformly continuous in ϕ, and the proposition follows.

The following example shows that it is not enough that $w_\phi(x)$ is uniformly continuous in ϕ for each $x \in \mathcal{X}$.

Example 3. Define $\pi(x) = 1/x^2$ for $x \in \mathcal{X} = [1, \infty)$. For $\phi \in \Phi = [0, 1/2]$ define $p_\phi(x) = c_\phi g_\phi(x)$, where $g_0(x) = \pi(x)$, $g_\phi(x) = (1/x^2)[I(|x-1/\phi| > 1)+I(|x-1/\phi| \leq 1)((1-\phi)|x-1/\phi|+\phi)], 0 < \phi \leq 1/2$. Then each p_ϕ is continuous. Clearly, the normalizing constant c_ϕ is continuous in ϕ. It decreases to 1 as ϕ decreases to 0, and satisfies $1 = c_0 \leq c_\phi \leq c_{1/2} < 1/(\int_3^\infty (1/x^2)dx) = 3$. Clearly, $w_\phi(x) = \pi(x)/p_\phi(x)$ is maximized for $x = 1/\phi$, giving $w_\phi^* = 1/(c_\phi \phi) > 1/(3\phi)$. Hence w_ϕ^* is not bounded. This occurs even though $w_\phi(x)$ is continuous in ϕ for each x, in fact uniformly continuous by the compactness of Φ.

Appendix B Convergence without boundedness of $w_\phi(x)$.

The easiest way to obtain convergence for an AIC algorithm defined by a family $\{P_\phi\}$ of proposal distributions for which $w_\phi(x)$ is not bounded, is to add to the proposal distribution a component with sufficiently heavy tails (cf. section 3). If the tails are heavier than those of Π, this works at least if the density π is bounded. In this appendix we show how convergence can be obtained by allowing the sequence lengths to vary and to depend on the history. Even though we can not describe the sequence lengths as explicit functions of the history, we think that this result tells an important part of the story about why the AIC algorithm seems to work in practice even if $w_\phi(x)$ is not bounded.

Suppose the basic defining ingredients for an AIC are given, i.e. we have

(i) An initial proposal distribution P_0.

(ii) Distribution characteristics $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_r)$, of the form $\tilde{\theta}(P) = E_P(\psi(X)) = (E_P(\psi_1(X)), \ldots, E_P(\psi_r(X))) \in \Theta$.

(iii) A family of proposal distributions $P_\phi, \phi \in \Phi$, and a continuous function $\tilde{\phi} : \Theta \to \Phi$

(iv) A criterion for fixing the proposal at P_{ϕ_M} as in (4).

In order to specify an AIC procedure, it only remains to add to (i) - (iv) the specification of sequence lengths. For the sake of stating and proving the theorem below, we must consider a version of the AIC where the proposal distribution is not fixed after the criterion (iv) is satisfied. Let the sequence lengths for this version be $K_1, K_2, \ldots$. These may vary from run to run of the algorithm. For the standard version, we fix the sequence length at K_{M+1} from the $(M+1)$-st sequence onwards. Define $L_m = K_1 + \cdots + K_m$ = the total number of iterations at the end of the m-th sequence. We denote by $\{Z^t\}$ the chain arising by not fixing the proposal at P_{ϕ_M} after the burn in, i.e. by letting the adaptation go on indefinitely. Then Z^t coincides with X^t for $t \leq L_{M+1}$, but differs later because the proposal distribution differs. Define

$$\theta^m = \hat{\theta}(Z^{L_{m-1}+1}, \ldots, Z^{L_m}) = (1/K_m) \sum_{t=L_{m-1}+1}^{L_m} \psi(Z^t), m = 1, 2, \ldots \tag{12}$$

and

$$\phi_m = \tilde{\phi}(\theta^m), m = 1, 2, \ldots \tag{13}$$

We then have the following result:

Theorem 2 *There exist sequence lengths $K_1, K_2, \ldots$, where K_m may depend on Z^t for $t \leq L_{m-1}$, such that*

(i) θ^m converges almost surely to $\theta_0 = \tilde{\theta}(\Pi)$

(ii) ϕ_m converges almost surely to $\phi_0 = \tilde{\phi}(\theta_0)$

(iii) M is finite almost surely.

(iv) The distribution P^t of X^t converges to Π in total variation norm.

Proof: Define $\theta^0 = \tilde{\theta}(P_0)$, the distribution characteristics of the initial distribution. We may assume without loss of generality that the Euclidean distance $|\theta^0 - \theta_0| \leq 1$. Since θ^1 is defined as a sample mean, it follows by the ergodicity of the Metropolis-Hastings algorithm that there exists K_1 such that $P(|\theta^1 - \theta_0| > 1) \leq 1/2$. In the same way, there exists K_2, which may depend on $Z^1, \ldots, Z^{K_1}$ through the proposal distribution parameter vector ϕ_1 and the initial value Z^{K_1} for a new Metropolis-Hastings chain, such that $P(|\theta^2 - \theta_0| > 1/2) \leq (1/2)^2$. Continuing inductively, there exists K_m, which may depend on the proposal distribution parameter vector ϕ_{m-1} for the m-th sequence and the initial value $Z^{L_{m-1}}$, such that $P(|\theta^m - \theta_0| > 1/m) \leq (1/2)^m, m = 3, 4, \ldots$. Define the events $B = (\theta^m$ does not converge to $\theta_0)$ and $A_m = (|\theta^m - \theta_0| > 1/m), m = 1, 2, \ldots$ We then have that $B \subseteq \cap_{N=1}^{\infty} \cup_{m=N}^{\infty} A_m$. Choosing sequence lengths in the above manner, we have $P(A_m) \leq (1/2)^m$, and it follows that $P(B) \leq \lim_{N\to\infty} \sum_{m=N}^{\infty} P(A_m) = 0$, proving (i). Part (ii) follows from (i) by the continuity of $\tilde{\phi}$, while (iii) follows immediately from (ii). To prove (iv), let $\epsilon > 0$ be arbitrary. It follows from (iii) that L_M is also almost surely finite. Therefore, there exists t_0 such that $P(L_M > t_0) \leq \epsilon$. For $t > t_0$ and $A \subseteq \mathcal{X}$ we have

$$\begin{aligned} |P^t(A) - \Pi(A)| &\leq P(L_M > t_0)|P^t(A|L_M > t_0) - \Pi(A)| \\ &+ P(L_M \leq t_0)|P^t(A|L_M \leq t_0) - \Pi(A)| \leq \epsilon + |P^t(A|L_M \leq t_0) - \Pi(A)| \end{aligned} \tag{14}$$

Denote by Q_ϕ^0 the distribution of X^{t_0} given that $L_M \leq t_0$ and $\phi_M = \phi$, where ϕ is any parameter vector in Φ. Denote by Q_ϕ^s the distribution of V^s, where $\{V^s\}$ is the chain generated by an independent chain Metropolis-Hastings algorithm with Q_ϕ^0 as initial distribution and P_ϕ as proposal distribution. We then have for $t > t_0$

$$P^t(\cdot|L_M \leq t_0, \phi_M = \phi) = Q_\phi^{t-t_0}(\cdot) \tag{15}$$

For any $\phi \in \Phi$, choose s_ϕ such that $s \geq s_\phi$ implies that for all $A \subseteq \mathcal{X}$ we have $|Q_\phi^s(A) - \Pi(A)| \leq \epsilon$. Such an s_ϕ exists by the convergence in total variation norm of the Metropolis-Hastings algorithm, and may e.g. be taken as the smallest integer with this property. Then $S = s_{\phi_M}$ may be considered as a random variable, being a deterministic function of ϕ_M, which is almost surely finite given that $L_M \leq t_0$. Choose s_0 such that $P(S > s_0|L_M \leq t_0) \leq \epsilon$, and put $\Phi_0 = \{\phi \in \Phi : s_\phi \leq s_0\}$. We then have for any $A \subseteq \mathcal{X}$

$$\phi \in \Phi_0 \text{ and } s \geq s_0 \Rightarrow |Q_\phi^s(A) - \Pi(A)| \leq \epsilon \tag{16}$$

Furthermore, since clearly $\phi_M \in \Phi_0$ if and only if $S \leq s_0$,

$$P(\phi_M \notin \Phi_0|L_M \leq t_0) \leq \epsilon \tag{17}$$

By (15), (16) and (17) it follows that for $t \geq t_0 + s_0$ and $A \subseteq \mathcal{X}$

$$\begin{aligned} |P^t(A|L_M \leq t_0) - \Pi(A)| &\leq P(\phi_M \in \Phi_0|L_M \leq t_0)|P^t(A|L_M \leq t_0, \phi_M \in \Phi_0) - \Pi(A)| \\ &+ P(\phi_M \notin \Phi_0|L_M \leq t_0)|P^t(A|L_M \leq t_0, \phi_M \notin \Phi_0) - \Pi(A)| \\ &\leq |P^t(A|L_M \leq t_0, \phi_M \in \Phi_0) - \Pi(A)| + P(\phi_M \notin \Phi_0|L_M \leq t_0) \leq 2\epsilon \end{aligned} \tag{18}$$

Combining (14) and (18) finally gives (iv).

Remark 3 *Note that the structure of the algorithm ensures that ϕ_m is gradually attracted to ϕ_0. Hence, the proof is in line with the heuristic argument of section 3 and the idea expressed in Remark 1, and indeed no uniform boundedness condition is used. It is likely that the sequence length K_m will have to increase with m in order to meet the increasing precision requirement $P(|\theta^m - \theta_0| > 1/m) \leq (1/2)^m$). On the other hand, the proposal distribution P_{ϕ_m} should be an ever improving approximation to Π, working against this increase in K_m at least for a while.*

Remark 4 *It is an obvious drawback of this result that the sequence lengths are not specified explicitly. However, the ordinary Metropolis - Hastings algorithm and indeed practically any MCMC suffers from a similar weakness. The number of iterations both in the burn in and afterwards must be determined experimentally, using some kind of diagnostic test. We feel convinced that similar empirical methods could be used to determine reasonable sequence lengths in the case of an AIC. The construction of such methods is beyond the scope of this paper, however, and we will not persue this problem here.*